Extreme Zero-Shot Learning for Extreme Text Classification

Anonymous Author(s)

ABSTRACT

The eXtreme Multi-label text Classification (XMC) problem concerns finding most relevant labels for an input text instance from a large label set. Many real-world applications can be formulated as XMC problems, such as recommendation systems, document tagging and semantic search. However, the XMC setup faces two challenges: (1) it is not generalizable to predict unseen labels in dynamic environments, and (2) it requires a large amount of supervised (instance, label) pairs, which can be difficult to obtain for emerging domains. Recently, Gupta et al. [17] studied the generalized zero-shot XMC (GZ-XMC) setup and propose ZestXML to handle the unseen labels, which still requires a large number of annotated (instance, label) pairs. In this paper, we consider a more practical scenario called Extreme Zero-Shot XMC (EZ-XMC), in which no supervision is needed and merely raw text of instances and labels are accessible. Few-Shot XMC (FS-XMC), an extension to EZ-XMC with limited supervision is also investigated. To learn the semantic embeddings of instances and labels with raw text, we propose to pre-train Transformer-based encoders with self-supervised contrastive losses. Specifically, we develop a pre-training method **MACLR**, which thoroughly leverages the raw text with techniques including **M**ulti-scale **A**daptive **C**lustering, **L**abel **R**egularization, and self-training with pseudo positive pairs. Experimental results on four public EZ-XMC datasets demonstrate that MACLR achieves superior performance compared to all other leading baseline methods, in particular with approximately 5-10% improvement in precision and recall on average. Moreover, we also show that our pre-trained encoder can be further improved on FS-XMC when there are a limited number of ground-truth positive pairs in training. By fine-tuning the encoder on such a few-shot subset, MACLR still outperforms other extreme classifiers significantly.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; *semantical embeddings*; *self-supervised learning*.

KEYWORDS

self-supervised learning, contrastive learning, zero-shot learning, extreme classification

ACM Reference Format:

Anonymous Author(s). 2021. Extreme Zero-Shot Learning for Extreme Text Classification. In *Woodstock '18: ACM Symposium on Neural Gaze Detection, June 03–05, 2018, Woodstock, NY*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/1122445.1122456

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Woodstock '18, June 03–05, 2018, Woodstock, NY

© 2021 Association for Computing Machinery.

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/10.1145/1122445.1122456

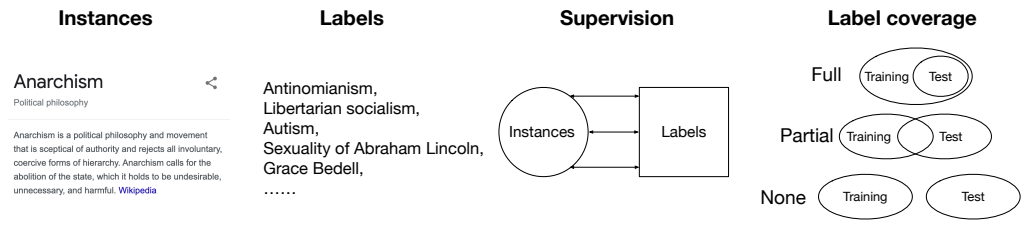
1 INTRODUCTION

The e**X**treme **M**ulti-label text **C**lassification (XMC) problem aims at tagging a text input with most relevant subset of labels from an extremely large output space. Many web-related applications can be formulated as an XMC task with encouraging results, such as finding the best matching products from a large catalog in e-commerce systems [6, 28], auto-completing queries given its prefix on search engines [41], predicting search keywords for dynamic advertising [8, 31], tagging categories of Wikipedia articles from a large label taxonomy [5, 12], to name just a few.

The current XMC setup is built on full label coverage and full supervision, where full label coverage means labels to be predicted have already appeared in the training set and full supervision indicates it requires a significant number of annotated (instance, label) pairs. In detail, it is assumed that an XMC algorithm has access to raw text of instances and labels, together with their corresponding relations during training, as shown in Figure 1.

However, there are several limitations of this XMC setting. First of all, due to the assumption of full label coverage, it is typical in XMC approaches to simply treat labels as IDs for classification and thus they are restricted to making predictions within observed labels. This assumption is unrealistic since the label set usually keeps growing over time, e.g., newly added websites or products which are absent during training yet crucial for applications such as recommendation and advertising. Besides, collecting labeled pairs is time-consuming, expensive and sometimes infeasible, for example, launching an e-commerce system in the emerging locale, where no user behavioral signals are avaiable. In spite of these constraints, most existing methods [10, 11, 29, 43] followed this XMC setup. It can be seen in Figure 2 that Astec [11], one of the state-of-the-art extreme classifiers, is incapable of handling the scenario without supervision, which leads to zero performance in both Precision@5 and Recall@100. Moreover, the increasing trend in Astec's performance along with the label ratio suggests that it depends highly on the supervision level and is hard to generalize to unseen labels. This motivates us to investigate how to design an effective XMC model with zero supervision.

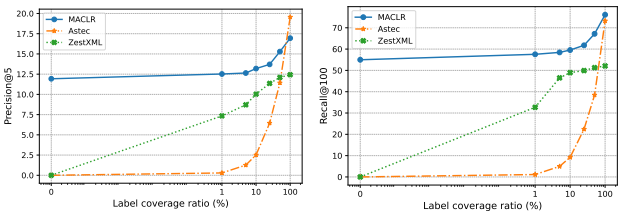
In this paper, we consider an essential yet under-explored XMC setting, called Extreme Zero-shot XMC (EZ-XMC). As depicted in Figure 1, we can access raw text of both instances and labels but do not know their corresponding relations in EZ-XMC. Moreover, we do not make any assumption on the label coverage, so the labels in the testing set may or may not appear in the training stage. An extension to EZ-XMC with a limited number of training pairs, Few-shot XMC (FS-XMC), is also taken into account in our paper. Either EZ-XMC or FS-XMC occurs frequently in the real world since informative and abundant (instance, label) pairs are never easy to obtain. Also, it is more practical and worthwhile to reduce labor for manual anotation by solving problems under EZ-XMC.



XMC	√	√	√	Full
GZ-XMC	√	√	√	Full & Partial & None
FS-XMC	√	√	√*	Full & Partial & None
EZ-XMC	√	√	×	Full & Partial & None

Figure 1: Four different settings in XMC. Four essential components are considered: instances (raw text), labels (raw text), supervision (positive pairs), and label coverage. In detail, we divide label coverage into 3 groups: full, partial, and none. * in FS-XMC emphasizes that only a limited amount of supervision is available. We can see that EZ-XMC is the most general and practical setting, where no supervision and label coverage is required.

For example, the unsupervised model can help narrow the tagging space remarkably by selecting a small number of candidate labels for products in the e-commerce domain for efficient annotation. Note that generalized zero-shot XMC (GZ-XMC) proposed in a recent work [17] can be regarded as a special case of EZ-XMC. GZ-XMC allows that the set of test labels is not completely overlapped with training labels but still requires supervision from positive pairs, as shown in Figure 1. From Figure 2, we can observe that ZestXML [17] designed for GZ-XMC also suffers the issue of no supervision although it can improve the precision and recall significantly under a low label coverage ratio compared with other standard XMC methods such as Astec [11].



(a) Precision@5. **(b) Recall@100.**

Figure 2: Performance of three representative XMC methods on the LF-Amazon-131K at different ratios of label coverage. A subset of (instance, label) pairs covering $[0, 1, 5, 10, 25, 50, 100](\%)$ of the whole label set is sampled.

A natural question then arises: how should we deal with EZ-XMC problems? Despite the name, EZ-XMC is barely easy to tackle. Fortunately, although dedicated supervision signals are lacking, raw text of instances and labels, e.g., product descriptions and categories, are still accessible in EZ-XMC. Thus it is of vital importance to effectively leverage self-information of these data to train a model for classification. To overcome challenges encountered in EZ-XMC, we turn to solving the problem from a different perspective without learning classifiers explicitly. In particular, XMC can be cast into a problem which learns a sentence encoder $\mathcal{E}$ to map instances and labels into dense embeddings, and predictions are made through approximate nearest neighbor search algorithms in the latent space [36]. Motivated by recent progresses in self-supervised learning [9, 13, 14, 20], we propose **MACLR** (**M**ulti-scale **A**daptive **C**lustering & **L**abel **R**egularization), a two-stage pre-training procedure with those unpaired raw data to obtain a sentence encoder $\mathcal{E}$ under EZ-XMC. Specifically, the Inverse Cloze Task (ICT) [24] is adopted to construct (context, title) pairs as the unsupervised anlog of (instance, label) pairs. To make full use of data's self-information, we leverage a multi-scale approach, which conducts clustering for instances with adaptive sizes along with the training process and treats those instances belonging to the same cluster as positive examples in contrastive learning. It allows the encoder to learn from the coarse to the fine scale gradually. Label information is utilized as a regularization term to separate an instance from its irrelevant labels. Moreover, self-training is conducted to further improve the encoder on pseudo positive pairs generated by the initial predictions of our model and TF-IDF. As to FS-XMC, fine-tuning the encoder on a few paired data is sufficient for the performance boost. Figure 2 demonstrates that MACLR achieves superior performance when no supervision is available and achieves much higher recall than Astec and ZestXML by a large margin, even under the higher label coverage ratio.

Our main contributions are summarized as follows:

- We propose an essential Extreme Zero-Shot XMC (EZ-XMC) setting without any assumptions on supervision and label coverage, which has not been explored in previous work and is more practical in real applications.
- We leverage unlabeled data to pretrain the sentence encoder $\mathcal{E}$ with improved Inverse Cloze Task in Stage I of MACLR. In particular, multi-scale adaptive clustering and label regularization are proposed to utilize raw text thoroughly. In

Stage II, we further self-train the encoder with pseudo positive pairs constructed from $\mathcal{E}$ in Stage I as well as TF-IDF model with complementary information.

- Comprehensive experiments are conducted on four public benchmark EZ-XMC datasets. Results demonstrate that our pre-trained encoder can outperform existing unsupervised baseline methods notably. As an example, MACLR achieves Recall@100 of 54.99%, nearly the same level as Astec (one of the SOTA XMC methods) [11] trained with a supervised subset covering around 70% labels on LF-Amazon-131K.
- MACLR can also achieve comparable or even better performance under the few-shot setting than those models heavily dependent on supervised information. For example, MACLR is better than the SOTA ZestXML [17] in Recall@100 over 20% (57.55% v.s. 32.69%) when fine-tuned on the subset covering 1% labels of LF-Amazon-131K.

2 RELATED WORK

Extreme multi-label classification. Various extreme classifiers have been proposed to address the large output space challenge of XMC problems. We can broadly categorize them into two groups: partitioned-based models with linear classifiers [31, 32, 44] that partition labels with hierarchical trees, leading to sub-linear inference time complexity, and embedding-based methods [4, 15, 21] that learn a classifier for each label and leverage approximated nearest neighbor [16, 27] to index labels in the large output space. There are also deep learning models such as AttentionXML [43], Astec [11], SiameseXML [10], and XR-Transformer [45] that further improve the accuracy of those linear counterparts with various advanced encoder architectures. Nevertheless, none of those XMC methods can handle the EZ-XMC setup: they not only suffer from the lack of supervised signals, but also fail to generalize to unseen cold-start labels in the test set.

The only exception is ZestXML [17], a recently proposed XMC method that was designed to address the generalized zero-shot XMC (GZ-XMC) problem where a number of labels for prediction are absent during training. While ZestXML partially resolves the generalization challenge of cold-start labels, just like those conventional XMC models, ZestXML still depends heavily on a large number of training data with positive (instance, label) pairs. None of existing methods consider the scenario where no supervision information is available. MACLR, on the other hand, can make satisfactory predictions even *without any paired data* and outperform all baseline methods when there is a limited amount of supervision.

Dense sentence embeddings. With increasinglg powerful computing resources, training dense sentence embeddings has become popular to learn the semantics directly from raw text [26, 35, 40]. Based on dense embeddings, approaches such as Sentence-BERT [35] and TwinBERT [26] project an instance and its relevant labels together in a shared latent space. Since the embeddings of all candidate labels can be precomputed and indexed, the label prediction can be done efficiently with approximate nearest neighbor search algorithms in the embedding space [16, 36]. Such methods can also be extended to applications with millions of labels, and can be utilized to produce meaningful embeddings for even unseen sentences due to the generalization of neural networks.

Self-supervised learning techniques. The past few years have witnessed great promise in self-supervised learning [9, 13, 14, 20, 22, 23], where a pre-training task is defined using only data's self-information. Learned representations from the pre-training task can be then leveraged in a wide range of downstream tasks in various domains, such as image classification [9, 20] and object detection [25] in computer vision, and open-domain question answering [18, 24] in natural language processing. Specifically, we focus on contrastive approaches for Sentence-BERT [35] models in this paper, where the intuition is to pull semantically close neighbors together and push apart non-neighbors via noise contrastive estimation or N-pair losses. Various effective pre-training tasks such as Inverse Cloze Task (ICT) [24] and SimCSE [14] have been shown to improve the performance of Sentence-BERT models.

3 PROBLEM FORMULATION

In this section, we present the problem formulation of EZ-XMC. As introduced before, with $\mathcal{X}$ and $\mathcal{Y}$ denoting the set of instances and labels respectively, the general XMC problem can be viewed as learning a scoring function $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. $f(\cdot, \cdot)$ maps an (instance, label) pair (x, y) to a similarity score, which is used to make a prediction through approximate nearest neighbor search algorithms. In previous settings such as XMC and GZ-XMC, a considerable amount of relevant (instance, label) pairs $\{(x_i, y_i)\}$ are available. On the contrary, in EZ-XMC, we have no knowledge about corresponding relations between instances and labels, but only their raw text, as shown in Figure 1. In this case, existing approaches that depend on the relevant pairs fail to learn an effective scoring function, even with a few paired data under FS-XMC.

Recent progresses in self-supervised learning have shown that a generalized sentence encoder can be learned through elaborately designed pre-training tasks even without any supervision [7, 24], and then adapted to different downstream tasks directly or via slight finetuning. On the other hand, the scoring function f can be modeled as

$$f(x, y) = \langle \mathcal{E}(x), \mathcal{E}(y) \rangle, \tag{1}$$

where $\mathcal{E}$ is a sentence encoder producing semantical dense embeddings, and $\langle \cdot, \cdot \rangle$ is the similarity measurement such as inner product and cosine similarity. Without loss of generality, inner product is adopted in the paper as the similarity metric between embeddings of instances and labels. Thus, we formulate the problem as training such an encoder $\mathcal{E}$ with raw text of $\mathcal{X}$ and $\mathcal{Y}$ through a pre-training task for EZ-XMC. As to the few-shot scenario FS-XMC where a limited number of pairs are available, we can fine-tune $\mathcal{E}$ for improvement. We will introduce the proposed algorithm for training $\mathcal{E}$ in Section 4.

4 METHOD

In this section, we introduce a two-stage pre-training procedure, MACLR, to thoroughly leverage unpaired data with raw text for EZ-XMC. Specifically, we present the general framework in Section 4.1, and then dive into details of two stages, pre-training with the improved Inverse Cloze Task and self-training with pseudo positive pairs, in Sections 4.2 and 4.3 respectively.

4.1 Framework

The framework of our pre-training procedure is shown in Figure 3. MACLR consists of two stages:

- Stage I: title-context pairs are constructed for the Inverse Cloze Task, and the encoder $\mathcal{E}$ is then trained on these pairs together with two proposed techniques, multi-scale adaptive clustering and label regularization.
- Stage II: More pseudo positive pairs are crafted using different score functions modeled by the encoder from Stage I and TF-IDF respectively. $\mathcal{E}$ is further trained on additional pairs to improve the encoding performance.

We will discuss the details of each component in our pre-training framework in the following sections.

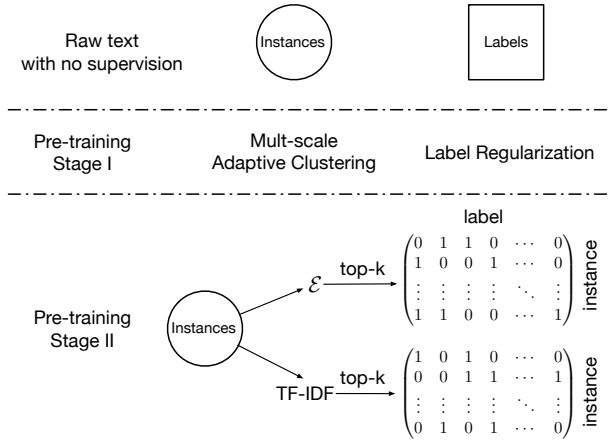


Figure 3: Framework of our pre-training procedure.

4.2 Stage I: Pre-training with the improved ICT

Inverse Cloze Task [24] is a frequently used pre-training task for the sentence encoder. Specifically, for an instance $x = \{s_1, \ldots, s_n\}$ consisting of n sentences, ICT randomly samples a sentence to serve as the pseudo positive label $\hat{y} = s_i$ where $i \sim [1, n]$. Then the rest of x is the pseudo instance $\hat{x} = \{s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n\}$. In XMC, due to the property that the label usually summarizes the instance with one short sentence, which works similarly as the title s_1, we directly utilize (context, title) pairs in the form of $(\hat{x} = \{s_2, \ldots, s_n\}, \hat{y} = s_1)$. This contruction is better to work as the analog of the ground truth (instance, label) pairs and capture the semantics of a sentence. With these pseudo pairs, the contrastive training objective for a mini-batch of N pairs is as follows:

$$\mathcal{L}_{\text{contrastive}} = -\sum_{i=1}^{N} \log \frac{\exp(\mathcal{E}(\hat{x}_i) \cdot \mathcal{E}(\hat{y}_i))}{\sum_{j=1}^{N} \exp(\mathcal{E}(\hat{x}_i) \cdot \mathcal{E}(\hat{y}_j))} \tag{2}$$

Based on ICT, we also develop two techniques, multi-scale adaptive clustering and label reguarization, to fully leverage the information of unpaired instances and labels.

4.2.1 Multi-scale Adaptive Clustering. In the original ICT scheme, we can construct only one positive pair for a particular instance. It is relatively hard in contrastive learning without enough positive examples, especially for extreme multi-label classification where one instance might be associated with more than one label, and a label is also likely to point to several different instances at the same time. Thus a question arises naturally: is it possible to construct more positive pairs from purely unpaired raw data to intergrate richer information into the pre-training process? We solve it by the unsupervised K-means clustering. In detail, we divide pseudo (context, title) pairs from ICT into K clusters through K-means based on the embeddings of all instances. Then if $C(\hat{x}_i) = C(\hat{x}_j)$, i.e., $\hat{x}_i$ and $\hat{x}_j$ belong to the same cluster, $(\hat{x}_i, \hat{y}_j)$ and $(\hat{x}_j, \hat{y}_i)$ are regarded as positive pairs besides original ICT pairs. Furthermore, supervised contrastive loss is adopted for training the encoder with a mini-batch of N pairs based on the cluster assignment:

$$\mathcal{L}_{\text{cluster}} = \sum_{i=1}^{N} \frac{-1}{|P_{\mathcal{Y}}(i)|} \sum_{p \in P_{\mathcal{Y}}(i)} \log \frac{\exp(\mathcal{E}(\hat{x}_i) \cdot \mathcal{E}(\hat{y}_p))}{\sum_{j=1}^{N} \exp(\mathcal{E}(\hat{x}_i) \cdot \mathcal{E}(\hat{y}_j))} \tag{3}$$

Here, $P_{\mathcal{Y}}(i) = \{p \in \{1, \ldots, N\} : C(\hat{x}_i) = C(\hat{x}_p)\}$ is the set of indices of all positives for $\hat{x}_i$ in the batch, and $|P_{\mathcal{Y}}(i)|$ is its cardinality. Minimizing Equation (3) pulls close the representations of instances and their positive labels within the same cluster and pushes away the representations of those from different clusters.

Besides, since the ultimate goal is the minimization of Equation (2), we propose a multi-scale approach with adaptive training, which guides the encoder to learn the easier tasks with sufficient positive examples, and then master harder tasks gradually. This approach allows the encoder to learn from the coarse scale to the fine scale of clustering assignment, and is similar to the idea of curriculum learning [2] to first focus on learning from a subset of simple examples, and expanding to include the remaining harder samples. Our adaptive training process can be conducted by modifying the cluster size to adjust the task difficulty accordingly. To be specific, we initialize the cluster assignment with the number of clusters K_0, and double the cluster size every T steps. The cluster assignment is also updated every T_{update} steps along with the training of $\mathcal{E}$. Such a process lasts for half of the total training steps T_{total} to take advantage of positive examples from constructed clusters. The obtained intermediate encoder from this adaptive procedure is expected to satisfactorily capture the semantics of a sentence and is ready to deal with the optimization with Equation (2). Then for the rest half of training steps, we turn to the hardest setting treating each instance as one independent cluster, which exactly falls into the contrastive training objective in Equation (2). Our multi-scale adaptive clustering is illustrated in Figure 4.

4.2.2 Label Regularization. In addition to leveraging information from the instance side, we also have access to the raw texts of the whole label set and can utilize them to boost the encoder's performance from the label side. Intuitively, for a randomly sampled label, with a high probability it is an negative example to the instance of interest. We can take advantage of this intuition to make the embedding of the instance far apart from its irrelevant labels. Instead of increasing the distance directly, it is more stable and effective to adopt contrastive losses. To avoid overfitting, we choose a new positive example for each instance instead of its corresponding pseudo label from ICT which has been used in $\mathcal{L}_{\text{cluster}}$. More concretely, $\hat{x}_i^+$ is selected exactly the same as $\hat{x}_i$, since the dropout

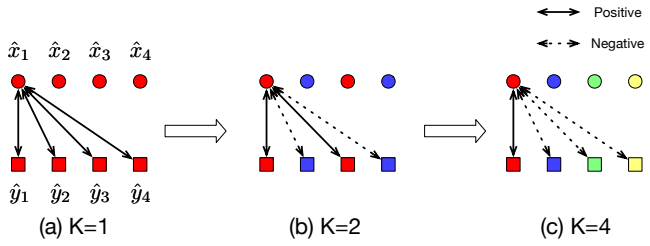


Figure 4: An example of multi-scale adaptive clustering. Here different colors represent different clusters. (a) In the beginning, there is only one cluster and $\{\hat{y}_j\}_{j=1}^4$ are all positive labels for $\hat{x}_1$. (b) K is doubled to 2 and now $\hat{y}_1$ and $\hat{y}_3$ are positive to $\hat{x}_1$. (c) Finally, K is equal to 4 where each instance itself is a cluster, and hence $\hat{x}_1$ only has one positive label $\hat{y}_1$. The process is similar for the rest of the instances.

layer is placed in the standard training of Transformer-based models and can be viewed as a minimal form of data augmentation [14]. By feeding the same sentence to the encoder $\mathcal{E}$, two embeddings with different dropout masks are obtained, i.e., $\hat{h}_i = \mathcal{E}(\hat{x}_i, z_i)$ and $\hat{h}_i^+ = \mathcal{E}(\hat{x}_i^+, z_i^+)$ where z represents a random mask for dropout. Note that $\hat{h}_i \neq \hat{h}_i^+$ due to the dropout noise, but they hold similar semantics from the same sentence and thus can be used as a positive pair for contrastive learning. The procedure of label regularization is depicted in Figure 5. At each training step, we sample M real labels from the label set $\mathcal{Y}$, and the reguarization term is computed as follows:

$$\mathcal{L}_{\text{label}} = \sum_{i=1}^{N} -\log \frac{\exp(\hat{h}_i \cdot \hat{h}_i^+)}{\exp(\hat{h}_i \cdot \hat{h}_i^+) + \sum_{j=1}^{M} \exp(\hat{h}_i \cdot \mathcal{E}(y_j^-))} \tag{4}$$

Through minimizing $\mathcal{L}_{\text{label}}$, the encoder learns to pull the instance away from its irrelevant labels and incorporate the dropout augmentation at the same time. Together with $\mathcal{L}_{\text{cluster}}$, we have the final objective function for pre-training in the Stage I as

$$\mathcal{L} = \mathcal{L}_{\text{cluster}} + \mathcal{L}_{\text{label}}. \tag{5}$$

We present the detailed implementation in Algorithm 1.

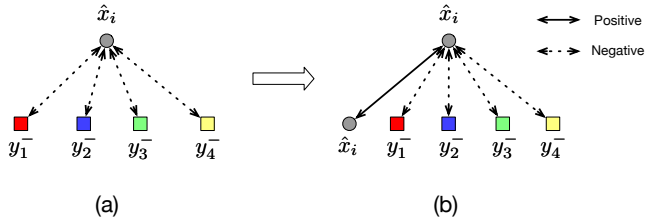


Figure 5: An illustration of label reguarization. (a) shows that $\hat{x}_i$ is expected to be far apart from sampled irrelevant labels $\{y_j^-\}_{j=1}^4$, while (b) indicates the identical $\hat{x}_i$ is added as a positive example for label regularization.

Algorithm 1 Pre-training procedure of MACLR

Input: Raw text of instances and labels $(\mathcal{X}, \mathcal{Y})$, the sentence encoder $\mathcal{E}$, batch size N and M, training step parameters T_K, T_{update} and T_{total}, initial cluster size K_0, # of top candidates k
Output: A pre-trained sentence encoder $\mathcal{E}$

▷ Stage I: Pre-training with the improved ICT
Construct ICT (context, title) pairs from raw texts in $\mathcal{X}$
Feed the context for each pair into the encoder $\mathcal{E}$ and cluster them into $K = K_0$ clusters via k-means
for $t = 1, \ldots, T_{\text{total}}$ **do**
  Sample a mini-batch of pseudo pairs of size N and a mini-batch of real labels of size M
  Compute the loss: $\mathcal{L} = \mathcal{L}_{\text{cluster}} + \mathcal{L}_{\text{label}}$
  Train the encoder by minimizing $\mathcal{L}$
  if t mod $T_K = 0$ and $t < T_{\text{total}}/2$ **then**
    $K = K * 2$
  end if
  if t mod $T_{\text{update}} = 0$ and $t < T_{\text{total}}/2$ **then**
    Feed raw texts of $\mathcal{X}$ again into $\mathcal{E}$, and update current cluster assignment via k-means with the cluster number K
  end if
  if $t \geq T_{\text{total}}/2$ **then**
    Treat each instance as an independent cluster
  end if
end for
▷ Stage II: Self-training with multi-viewed pseudo pairs
Construct pseudo pairs $(\mathcal{X}_{\text{pseu}}, \mathcal{Y}_{\text{pseu}})$ by selecting top-k candidate labels with the similarity metric in Equation (1) on the encoder $\mathcal{E}$ and TF-IDF respectively
Train the encoder $\mathcal{E}$ for T_{total} steps by minimizing Equation (3)

4.3 Stage II: Self-training with multi-viewed pseudo pairs

After the pre-training procedure in Section 4.2, we can obtain an intermediate encoder $\mathcal{E}_I$. But are there any ways to further improve the encoder? On the one hand, inspired by self-training in semi-supervised learning [19, 39, 42, 46], $\mathcal{E}_I$ can be leveraged to make predictions on those unpaired training instances themselves, to generate pseudo positive pairs. These pseudo pairs are much better than random guessing and can serve as a distinct view from ICT pairs. On the other hand, similar pseudo pairs can be constructed by other unsupervised methods such as TF-IDF, which also provide different and complementary information about the instance.

With multi-viewed pseudo positive pairs, we can conduct further training on the encoder in State II from a new perspective and self-improve $\mathcal{E}_I$. The detailed process works as follows:

1) Compute the similarity score in Equation (1) using $\mathcal{E}_I$ for each training instance i, and select labels with top-k maximum scores as its pseudo labels;
2) Generate labels similarly with TF-IDF, except that $\mathcal{E}(x)$ and $\mathcal{E}(y)$ are replaced with their TF-IDF vectors;
3) Mix pseudo positive pairs from 1) and 2) together, and train $\mathcal{E}_I$ on them with Equation (3). Here the positive index set for the instance $\hat{x}_i$ is defined as $P_{\mathcal{Y}}(i) = \{p | \hat{x}_p = \hat{x}_i\}$.

The whole pre-training procedure of MACLR is shown in Algorithm 1. Note that for FS-XMC, we simply fine-tune the encoder $\mathcal{E}$ from MACLR on available positive pairs for several steps by minimizing the original contrastive loss in Equation (2).

5 EXPERIMENTAL RESULTS

5.1 Experimental Settings

Datasets. We evaluate our proposed MACLR on 4 public XMC benchmark datasets [3, 17] where raw text of instances and labels are available. These datasets are derived from real-world applications, ranging from item-to-item recommendation (LF-Amazon-131K, LF-Amazon-1M), to Wikipedia articles category/title tagging (LF-WikiSeeAlso-320K, LF-Wikipedia-500K). All these datasets are available in the Extreme Classification Repository [3][1], except for LF-Amazon-1M, which is available from the ZestXML paper [17][2]. Detailed dataset statistics are presented in Table 1.

Table 1: Dataset statistics. N_{train}, N_{test} and N_{label} are the number of training points, test points, and labels respectively. D_{BoW} is the dimensionality of Bag-of-Words (BoW) features.

Dataset	N_{train}	N_{test}	N_{label}	D_{BoW}
LF-Amazon-131K	294,805	134,835	131,073	80,000
LF-WikiSeeAlso-320K	693,082	177,515	312,330	80,000
LF-Wikipedia-500K	1,813,391	783,743	501,070	500,000
LF-Amazon-1M	914,179	1,465,767	960,106	1,000,000

Evaluation Protocol. We consider two evaluation setups: Extreme Zero-shot Learning of XMC (EZ-XMC) and Few-shot Learning of XMC (FS-XMC). EZ-XMC is a fully unsupervised learning setup where no positive (instance, label) pairs are available. The only available information is the raw text of training instances and the whole label set. FS-XMC is a semi-supervised learning setup where only very few positive (instance, label) pairs in the training set are available. Regardless of the learning procedure, all models are evaluated on the same test set for fair comparison.

We evaluate the models' performance with precision@k (P@k, $k \in \{1, 3, 5\}$) and recall@k (R@k, $k \in \{1, 3, 5, 10, 100\}$), which are two commonly-used evaluation metrics in the XMC literature [6, 34]. In detail, $P@k$ and $R@k$ are defined as follows:

$$P@k = \frac{1}{k} \sum_{i \in \text{rank}_k(y')} y_i, \quad R@k = \frac{1}{\sum_l y_l} \sum_{i \in \text{rank}_k(y')} y_i, \qquad (6)$$

where $y \in \{0, 1\}^L$ and $y' \in R^L$ are the ground true vector and the prediction vector respectively. Here rank_k returns the indices of the top-k highest elements.

Baseline Methods. For the setting of EZ-XMC, we compare our method with the following unsupervised learning algorithms:

- TF-IDF [33], which represents instances and labels by sparse TF-IDF features and retrieves top labels for each instance based on the similarity of TF-IDF features;
- XR-Linear [44], a hierarchical linear model trained with pseudo positive pairs constructed from TF-IDF;
- GloVe [30], which adopts dense average word embeddings with the dimension of 300 trained on co-occurrence statistics to measure similarity between instances and labels;
- Sentence-BERT (SentBERT) [13, 35], a sentence encoder modeled as a Siamese-Transformer to derive semantically meaningful embeddings for instances and labels;
- Paraphrase MPNet (MPNet) [37], another Sentence-BERT model originally designed for searching sentence paraphrases;
- SimCSE [14], a Siamese-Transformer pre-trained with the contrastive objective using dropout noise as augmentation;
- ICT [24], another Siamese-Transformer pre-trained with the contrastive objective using (context, title) pairs.

Note that SentBERT and MPNet are pre-trained on external multi-task learning datasets with extra supervision. In contrast, SimCSE and ICT are fully unsupervised pre-rained Siamese-Transformers on the specific XMC dataset only.

For the setting of FS-XMC, as few-shot (instance, label) pairs are available, we additionally compare fine-tuned MACLR with competitive XMC approaches, including Astec [11], SiameseXML [10], and ZestXML [17]. It should be noticed that ZestXML is the leading XMC method that improves performance on few-shot labels. We

Table 2: Extreme Zero-shot Learning (EZ-XMC) comparison of different unsupervised methods.

Method	Precision			Recall				
	@1	@3	@5	@1	@3	@5	@10	@100
LF-Amazon-131K								
TF-IDF	12.38	11.50	9.14	6.91	18.14	23.21	29.32	45.04
XR-Linear	7.56	7.84	7.30	4.05	12.11	18.32	29.17	40.39
GloVe	3.67	2.78	2.15	2.05	4.33	5.44	7.23	14.17
SentBERT	1.86	1.44	1.14	1.01	2.22	2.88	4.01	10.18
MPNet	13.94	11.41	8.82	7.82	18.08	22.58	27.91	43.39
SimCSE	10.13	8.61	6.69	5.61	13.39	16.84	21.27	35.81
ICT	13.82	11.41	8.90	7.76	18.09	22.80	28.94	47.40
MACLR (ours)	**18.13**	**15.42**	**11.93**	**10.35**	**24.45**	**30.43**	**37.28**	**54.99**
LF-WikiSeeAlso-320K								
TF-IDF	10.71	8.90	7.15	5.92	13.03	16.48	21.60	42.55
XR-Linear	4.73	4.27	3.90	2.23	5.83	8.64	14.18	36.93
GloVe	3.86	2.76	2.21	2.12	4.11	5.22	6.95	15.33
SentBERT	1.71	1.27	1.06	1.08	2.16	2.90	4.17	10.76
MPNet	13.75	11.93	9.58	8.14	17.77	22.21	28.11	45.91
SimCSE	9.03	6.64	5.22	4.99	9.89	12.34	15.93	30.11
ICT	10.76	10.05	8.12	6.12	14.32	18.05	23.01	39.77
MACLR (ours)	**16.31**	**13.53**	**10.78**	**9.71**	**20.39**	**25.37**	**32.05**	**53.83**
LF-Wikipedia-500K								
TF-IDF	20.30	12.98	9.96	7.25	12.91	15.98	20.31	38.16
XR-Linear	10.67	8.77	7.61	3.69	8.58	12.11	19.80	31.02
GloVe	2.19	1.52	1.23	0.85	1.66	2.18	3.10	8.52
SentBERT	0.17	0.15	0.13	0.05	0.13	0.18	0.30	1.29
MPNet	22.46	12.87	9.49	8.74	14.07	16.76	20.64	34.72
SimCSE	14.32	6.84	4.55	4.24	8.03	11.26	14.35	27.68
ICT	17.74	9.67	7.06	7.35	11.60	13.84	17.19	31.08
MACLR (ours)	**28.44**	**17.75**	**13.53**	**10.40**	**18.16**	**22.38**	**28.52**	**50.09**
LF-Amazon-1M								
TF-IDF	7.68	9.20	7.23	5.61	19.30	24.92	31.76	51.79
XR-Linear	5.19	5.48	5.26	3.63	11.30	17.94	31.18	43.79
GloVe	4.05	4.07	3.07	2.91	8.42	10.44	12.90	21.18
SentBERT	2.82	2.87	2.13	2.03	5.91	7.21	8.80	14.22
MPNet	8.29	8.87	6.80	6.04	18.64	23.51	29.35	46.15
SimCSE	3.33	3.69	2.74	2.38	7.66	9.38	11.43	18.54
ICT	8.66	9.26	7.13	6.30	19.45	24.60	30.73	48.42
MACLR (ours)	**9.58**	**10.41**	**8.03**	**7.38**	**22.01**	**27.72**	**34.48**	**55.23**

[1] http://manikvarma.org/downloads/XC/XMLRepository.html
[2] https://github.com/nilesh2797/zestxml

Table 3: Results of few-shot learning XMC (FS-XMC) where the training subset covers 1% labels from the whole set.

Method	Precision			Recall				
	@1	@5	@10	@1	@3	@5	@10	@100
LF-Amazon-131K								
XR-Linear	1.53	0.57	0.36	0.67	0.75	0.78	0.81	0.92
Astec	0.94	0.44	0.29	0.55	0.78	0.84	0.91	1.13
SiameseXML	1.45	0.56	0.35	0.84	0.96	1.00	1.03	1.16
ZestXML	10.10	9.19	7.34	5.63	14.46	18.61	23.73	32.69
SentBERT	12.64	9.82	7.80	6.97	15.34	19.74	25.33	43.53
MPNet	14.78	11.55	8.97	8.28	18.24	22.84	28.54	45.89
MACLR (ours)	**18.74**	**16.07**	**12.52**	**10.73**	**25.44**	**31.89**	**39.17**	**57.55**
LF-WikiSeeAlso-320K								
XR-Linear	1.24	0.57	0.37	0.42	0.58	0.63	0.68	0.76
Astec	1.25	0.60	0.41	0.69	0.98	1.11	1.27	1.56
SiameseXML	1.81	0.75	0.48	1.03	1.26	1.33	1.41	1.67
ZestXML	8.74	6.78	5.41	4.68	9.70	12.21	15.73	24.98
SentBERT	16.30	12.62	10.08	9.30	18.92	23.78	30.40	52.92
MPNet	17.14	12.64	9.96	9.97	18.98	23.45	29.67	50.75
MACLR (ours)	**19.09**	**14.57**	**11.53**	**11.39**	**22.34**	**27.63**	**34.81**	**57.92**
LF-Wikipedia-500K								
XR-Linear	2.95	1.19	0.75	0.62	0.74	0.76	0.79	0.84
Astec	2.85	1.16	0.73	1.46	1.75	1.84	1.92	2.08
SiameseXML	2.72	1.15	0.73	1.39	1.73	1.84	1.93	2.09
ZestXML	23.86	14.97	11.31	7.19	13.00	16.03	20.13	29.95
SentBERT	32.09	20.50	15.78	10.94	19.46	24.12	30.94	55.94
MPNet	34.58	22.02	16.86	11.96	21.32	26.30	33.53	57.78
MACLR (ours)	**44.27**	**28.46**	**21.83**	**15.14**	**27.04**	**33.33**	**42.03**	**67.95**
LF-Amazon-1M								
XR-Linear	0.51	0.20	0.12	0.36	0.42	0.43	0.45	0.49
Astec	0.49	0.59	0.12	0.34	0.40	0.42	0.44	0.49
SiameseXML	0.60	0.73	0.15	0.41	0.46	0.48	0.49	0.53
ZestXML	5.07	5.89	4.38	3.68	12.31	15.04	17.80	22.51
SentBERT	6.56	6.93	5.68	4.35	18.29	24.72	28.69	48.52
MPNet	8.87	10.34	7.56	6.78	20.11	26.14	31.98	50.48
MACLR (ours)	**10.37**	**11.23**	**8.58**	**7.57**	**23.55**	**29.60**	**36.71**	**56.44**

Table 4: Results of few-shot learning XMC (FS-XMC) where the training subset covers 5% labels from the whole set.

Method	Precision			Recall				
	@1	@5	@10	@1	@3	@5	@10	@100
LF-Amazon-131K								
XR-Linear	5.09	2.09	1.32	2.36	2.86	3.02	3.18	3.74
Astec	3.94	1.92	1.26	2.31	3.34	3.66	4.00	4.96
SiameseXML	5.36	2.23	1.41	3.15	3.89	4.08	4.27	4.82
ZestXML	12.33	10.99	8.71	6.84	17.19	21.97	28.10	46.49
SentBERT	15.47	12.24	9.64	8.63	19.23	24.40	30.82	49.22
MPNet	15.03	11.88	9.28	8.47	18.74	23.69	29.93	48.84
MACLR (ours)	**19.56**	**16.19**	**12.64**	**11.15**	**25.65**	**32.18**	**39.63**	**58.45**
LF-WikiSeeAlso-320K								
XR-Linear	4.69	2.20	1.46	1.82	2.41	2.63	2.82	3.42
Astec	5.90	2.80	1.86	3.26	4.49	4.95	5.49	6.83
SiameseXML	6.83	3.15	2.06	3.88	5.15	5.56	6.02	7.09
ZestXML	10.06	8.11	6.60	5.33	11.49	14.74	19.57	40.46
SentBERT	18.47	14.19	11.29	10.82	21.55	26.77	33.92	57.02
MPNet	18.59	13.99	11.08	10.89	21.12	26.10	32.82	54.70
MACLR (ours)	**20.99**	**15.57**	**12.26**	**12.59**	**23.94**	**29.41**	**36.78**	**59.81**
LF-Wikipedia-500K								
XR-Linear	11.80	5.30	3.39	2.76	3.47	3.65	3.82	4.09
Astec	11.23	5.27	3.48	5.46	7.47	8.16	8.90	10.35
SiameseXML	12.44	5.69	3.79	6.05	7.98	8.62	9.22	10.40
ZestXML	27.31	17.31	13.09	8.28	15.13	18.64	23.30	36.50
SentBERT	41.06	26.35	20.25	14.17	25.34	31.32	39.77	66.24
MPNet	42.81	28.07	21.66	14.67	26.81	33.24	42.28	67.76
MACLR (ours)	**47.25**	**30.57**	**23.54**	**16.20**	**29.01**	**35.81**	**45.13**	**71.35**
LF-Amazon-1M								
XR-Linear	2.11	0.84	0.53	1.45	1.74	1.81	1.88	2.04
Astec	2.22	2.56	0.71	1.54	1.91	2.03	2.16	2.41
SiameseXML	2.60	3.01	1.06	1.81	2.20	2.30	2.41	2.60
ZestXML	7.17	8.35	6.36	5.18	17.49	21.88	26.80	36.51
SentBERT	8.89	10.02	7.93	7.00	21.58	27.35	33.98	54.28
MPNet	9.25	10.41	8.00	7.11	21.87	27.64	34.61	54.72
MACLR (ours)	**10.60**	**11.47**	**8.80**	**7.89**	**24.14**	**30.44**	**37.95**	**58.45**

also take into account of SentBERT [35] with further fine-tuning to demonstrate the effectiveness of our pre-training procedure.

Hyper-parameters. We use a Siamese Transformer model to embed both instances and labels. The encoder consists of a 12 layers BERT-base model, topped with a linear head projecting hidden state of the [CLS] token into a 512-dimensional embedding. The sequence length of the instance and the label is set to be 288 and 64 respectively. We pre-train the model on eight V100 GPUs for 100,000 steps with an Adam optimizer and batch size of 32 per GPU in both Stage I and Stage II. This pre-training process takes about 1 day. We adopt an initial learning rate 1×10^{-5} with the warm-up ratio 0.1, followed by a linear learning rate decay. For fine-tuning, the learning rate of Adam is set to 5×10^{-6} with 2000 training steps for the 1% label ratio and 10K training steps for the 5% label ratio. In the Stage I, we use the initial cluster size $K = 2048$ and set $T_K = 10000$ and $T_{\text{update}} = 5000$. As to the Stage II, top 3 ranked labels from predictions of the encoder and TF-IDF are selected to constitute the pseudo set for self-training.

For hyper-parameters of all baselines, we follow their default setups. All experiments are conducted on the AWS p3dn.24xlarge instance, consisting of 96 Intenl Xeon CPUs with 768 GB of RAM and 8 Nvidia V100 GPUs with 32 GB of memory each.

5.2 Zero-Shot Learning

In this section, we focus on extreme zero-shot learning (EZ-XMC), where no real positive (instance, label) pairs are accessible. Table 2 presents detailed performance of precision and recall on all four datasets. Our proposed MACLR consistently outperforms all comparing baselines with a large margin for all four datasets. Compared to the leading sparse method TF-IDF, MACLR has an average of 5.3% and 9.1% absolute improvement in Precision@1 and Recall@100, respectively. Compared to the leading neural model MPNet, MACLR has an average of 3.5% and 10.9% absolute improvement in Precision@1 and Recall@100, respectively.

Speaking of sparse lexical matching approaches, TF-IDF remains a tough-to-beat unsupervised baseline. Specifically, TF-IDF performs better than many BERT variants (e.g., SentBERT, SimCSE, ICT), which is aligned with the finding in recent zero-shot dense retrieval literature [1, 38]. This suggests the importance of designing proper self-supervised learning tasks for Transformer models in unsupervised EZ-XMC setup. Also note that XR-Linear is based on TF-IDF vectors whereas the noise from pseudo pairs makes it even inferior to the original TF-IDF method.

As for pre-trained SentBERT models, on the other hand, only MPNet shows comparable performance with TF-IDF. MPNet remains competitive because it was trained on a large supervised

Table 5: Ablation study on LF-Amazon-131K.

Index	Ablation Configuration				Precision			Recall				
	MAC *	LR *	$\mathcal{E}$ †	TFIDF †	@1	@3	@5	@1	@3	@5	@10	@100
1	No	No	No	No	13.82	11.41	8.90	7.76	18.09	22.80	28.94	47.40
2	Yes	No	No	No	15.79	13.16	10.22	8.85	20.90	26.27	32.61	49.83
3	No	Yes	No	No	16.02	13.29	10.28	9.04	21.27	26.51	32.97	50.34
4	Yes	Yes	No	No	16.37	13.71	10.65	9.29	21.63	27.03	33.93	51.45
5	Yes	Yes	Yes	No	17.01	14.75	11.41	9.72	23.33	29.04	35.20	53.55
6	Yes	Yes	No	Yes	16.51	14.12	10.92	9.52	22.43	28.02	34.64	52.78
7	Yes	Yes	Yes	Yes	18.13	15.42	11.93	10.35	24.45	30.43	37.28	54.99

* MAC represents adaptive clustering while LR stands for label regularization.
† Pseudo positive pairs are constructed from $\mathcal{E}$ or TFIDF.

corpus (out-of-domain) to learn semantics between paraphrasing sentences. Thus, MPNet should be viewed as a multi-task learning baseline with extra supervision. However, MACLR is significantly better than MPNet with an average improvement of 3.5% in P@1 and over 10% in R@100. Furthermore, MACLR also outperforms its counterparts which are trained with effective pre-training tasks such as SimCSE and ICT on the target dataset, showing the effectiveness of pre-training strategies like multi-scale adaptive clustering in MACLR. Overall, results in Table 2 demonstrates that MACLR is capable to learn informative embeddings and to make useful predictions even with no supervision. We will investigate each component in MACLR in Section 5.4 thoroughly.

5.3 Few-Shot Learning

We further conduct few-shot learning (FS-XMC) experiments in which different learning algorithms can access a limited number of positive (instance, label) pairs. To simulate the scenario of few-shot learning, we first manually sample a small ratio of labels, then collect all their positive instances from the training set as the final subset of positive (instance, label) pairs for model training. Results of FS-XMC methods fine-tuned with 1% and 5% labels are shown in Tables 3 and 4 respectively.

Our proposed MACLR outperforms all other baselines significantly, including variants of Siamese-Transformer models (e.g., SentBERT, MPNet) and major competitive XMC methods (e.g., XR-Linear, Astec and SiameseXML), on all four datasets. Note that SiameseXML is the state-of-the-art XMC method under the full supervision setup of XMC. Here, we again witness that existing XMC methods heavily rely on the supervision level as well as the full-coverage of label space for test set. MACLR, in contrast, still performs robustly under the FS-XMC setup, which enjoy larger applicability to emerging domains with many cold-start labels.

Crucially, even ZestXML tailored to address the challenging scenario of unseen labels cannot match the performance of MACLR. In particular, when focusing on the few-shot scenario with only 1% sampled labels, MACLR achieves 18.74% in P@1, improving the performance of Astec with 0.94% and ZestXML with 10.10% significantly. Besides, MACLR outperforms all other Sentence-BERT counterparts, validating the effectiveness of our pre-training procedure. As to fine-tuning on the subset with 5% labels, performance of all compared methods are improved as expected with more supervision. The relative rank among these methods remains the same, with MACLR still performing the best in terms of precision and recall on all four datasets.

5.4 Ablation Study

In this part, we conduct an ablation study to investigate each component in our pre-training procedure, including multi-scale adaptive clustering, label regularization, and self-training with pseudo positive pairs constructed from the encoder or TF-IDF. We add a component once a time on LF-Amazon-131K to observe its independent influence on the model performance. Table 5 presents detailed performance on seven different configurations.

For two techniques multi-scale adaptive clustering and label regularization during the Stage I, they can improve the performance of the encoder separately, as shown in the performance gain of the index 2 and 3 over the index 1. When combined, they can further improve the accuracy of the model, from 8.90% to 10.65% in P@5 and from 47.40% to 51.45% in R@100. As to the second stage, we explore the impact of self-training with pseudo positive pairs either from the encoder itself or TF-IDF. We can see from Table 5 that pairs from both $\mathcal{E}$ and TF-IDF contribute to the improvement in precision and recall gain over the index 5. It further validates that the encoder and TF-IDF provides complementary perspective when constructing pseudo positive pairs.

6 CONCLUSIONS

This paper is the first to investigate the problem of Extreme zero-shot XMC without any supervision. We develop a two-stage pre-training procedure MACLR to train a Sentence-BERT style encoder on pseudo (context, title) pairs constructed from raw text. We demonstrate that techniques including multi-scale adaptive clustering, label regularization and self-training contribute to the performance gain of the pre-trained encoder. In particular, MACLR outperforms all unsupervised baselines significantly when there are no (instance, label) pairs provided. It also offers leading accuracy in both precision and recall after fine-tuning on a limited number of paired data, compared with state-of-the art extreme classifiers. A future direction could be adding a ranker model after the encoder to improve performance on head labels.

REFERENCES

[1] Anonymous. 2022. Contrastive Pre-training for Zero-Shot Information Retrieval. In *Submitted to The Tenth International Conference on Learning Representations*. https://openreview.net/forum?id=c7S4WIlmu5 under review.

[2] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*. 41–48.

[3] K. Bhatia, K. Dahiya, H. Jain, P. Kar, A. Mittal, Y. Prabhu, and M. Varma. 2016. The extreme classification repository: Multi-label datasets and code. http://manikvarma.org/downloads/XC/XMLRepository.html

[4] Kush Bhatia, Himanshu Jain, Purushottam Kar, Manik Varma, and Prateek Jain. 2015. Sparse Local Embeddings for Extreme Multi-label Classification.. In *NIPS*, Vol. 29. 730–738.

[5] Ilias Chalkidis, Emmanouil Fergadiotis, Prodromos Malakasiotis, and Ion Androutsopoulos. 2019. Large-Scale Multi-Label Text Classification on EU Legislation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 6314–6322. https://doi.org/10.18653/v1/P19-1636

[6] Wei-Cheng Chang, Daniel Jiang, Hsiang-Fu Yu, Choon Hui Teo, Jiong Zhang, Kai Zhong, Kedarnath Kolluri, Qie Hu, Nikhil Shandilya, Vyacheslav Ievgrafov, et al. 2021. Extreme multi-label learning for semantic matching in product search. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 2643–2651.

[7] Wei-Cheng Chang, Felix X Yu, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar. 2020. Pre-training tasks for embedding-based large-scale retrieval. In *International Conference on Learning Representations*.

[8] Wei-Cheng Chang, Hsiang-Fu Yu, Kai Zhong, Yiming Yang, and Inderjit S Dhillon. 2020. Taming pretrained transformers for extreme multi-label text classification. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 3163–3171.

[9] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PMLR, 1597–1607.

[10] Kunal Dahiya, Ananye Agarwal, Deepak Saini, K Gururaj, Jian Jiao, Amit Singh, Sumeet Agarwal, Purushottam Kar, and Manik Varma. 2021. SiameseXML: Siamese Networks meet Extreme Classifiers with 100M Labels. In *International Conference on Machine Learning*. PMLR, 2330–2340.

[11] Kunal Dahiya, Deepak Saini, Anshul Mittal, Ankush Shaw, Kushal Dave, Akshay Soni, Himanshu Jain, Sumeet Agarwal, and Manik Varma. 2021. DeepXML: A Deep Extreme Multi-Label Learning Framework Applied to Short Text Documents. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 31–39.

[12] Ofer Dekel and Ohad Shamir. 2010. Multiclass-multilabel classification with more classes than examples. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. JMLR Workshop and Conference Proceedings, 137–144.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[14] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

[15] Chuan Guo, Ali Mousavi, Xiang Wu, Dan Holtmann-Rice, Satyen Kale, Sashank Reddi, and Sanjiv Kumar. 2019. Breaking the glass ceiling for embedding-based classifiers for large output spaces. In *Advances in Neural Information Processing Systems*.

[16] Ruiqi Guo, Sanjiv Kumar, Krzysztof Choromanski, and David Simcha. 2016. Quantization based fast inner product search. In *Artificial Intelligence and Statistics*. PMLR, 482–490.

[17] Nilesh Gupta, Sakina Bohra, Yashoteja Prabhu, Saurabh Purohit, and Manik Varma. 2021. Generalized Zero-Shot Extreme Multi-label Learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 527–535.

[18] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909* (2020).

[19] Junxian He, Jiatao Gu, Jiajun Shen, and Marc'Aurelio Ranzato. 2019. Revisiting self-training for neural sequence generation. *arXiv preprint arXiv:1909.13788* (2019).

[20] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9729–9738.

[21] Himanshu Jain, Venkatesh Balasubramanian, Bhanu Chunduri, and Manik Varma. 2019. Slice: Scalable linear extreme classifiers trained on 100 million labels for related searches. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 528–536.

[22] Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*.

[23] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

[24] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent Retrieval for Weakly Supervised Open Domain Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Florence, Italy, 6086–6096. https://doi.org/10.18653/v1/P19-1612

[25] Yandong Li, Di Huang, Danfeng Qin, Liqiang Wang, and Boqing Gong. 2020. Improving object detection with selective self-supervised self-training. In *European Conference on Computer Vision*. Springer, 589–607.

[26] Wenhao Lu, Jian Jiao, and Ruofei Zhang. 2020. TwinBERT: Distilling knowledge to twin-structured compressed BERT models for large-scale retrieval. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2645–2652.

[27] Yu A Malkov and Dmitry A Yashunin. 2018. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence* 42, 4 (2018), 824–836.

[28] Tharun Medini, Qixuan Huang, Yiqiu Wang, Vijai Mohan, and Anshumali Shrivastava. 2019. Extreme classification in log memory using count-min sketch: A case study of amazon search with 50m products. In *Advances in Neural Information Processing Systems*.

[29] Anshul Mittal, Noveen Sachdeva, Sheshansh Agrawal, Sumeet Agarwal, Purushottam Kar, and Manik Varma. 2021. ECLARE: Extreme Classification with Label Graph Correlations. In *Proceedings of the Web Conference 2021*. 3721–3732.

[30] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[31] Yashoteja Prabhu, Anil Kag, Shrutendra Harsola, Rahul Agrawal, and Manik Varma. 2018. Parabel: Partitioned label trees for extreme classification with application to dynamic search advertising. In *Proceedings of the 2018 World Wide Web Conference*. 993–1002.

[32] Yashoteja Prabhu and Manik Varma. 2014. FastXML: A fast, accurate and stable tree-classifier for extreme multi-label learning. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. 263–272.

[33] Anand Rajaraman and Jeffrey David Ullman. 2011. *Mining of massive datasets*. Cambridge University Press.

[34] Sashank J Reddi, Satyen Kale, Felix Yu, Daniel Holtmann-Rice, Jiecao Chen, and Sanjiv Kumar. 2019. Stochastic negative mining for learning with large output spaces. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 1940–1949.

[35] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992. https://doi.org/10.18653/v1/D19-1410

[36] Anshumali Shrivastava and Ping Li. 2014. Asymmetric LSH (ALSH) for sublinear time maximum inner product search (MIPS). In *Advances in Neural Information Processing Systems*.

[37] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. MPNet: Masked and permuted pre-training for language understanding. In *Advances in Neural Information Processing Systems*.

[38] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A Heterogenous Benchmark for Zero-shot Evaluation of Information Retrieval Models. *arXiv preprint arXiv:2104.08663* (2021).

[39] Qizhe Xie, Minh-Thang Luong, Eduard Hovy, and Quoc V Le. 2020. Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10687–10698.

[40] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2021. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *International Conference on Learning Representations*.

[41] Nishant Yadav, Rajat Sen, Daniel N. Hill, Arya Mazumdar, and Inderjit S. Dhillon. 2021. Session-Aware Query Auto-Completion Using Extreme Multi-Label Ranking. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (Virtual Event, Singapore) *(KDD '21)*. Association for Computing Machinery, New York, NY, USA, 3835–3844. https://doi.org/10.1145/3447548.3467087

[42] I Zeki Yalniz, Hervé Jégou, Kan Chen, Manohar Paluri, and Dhruv Mahajan. 2019. Billion-scale semi-supervised learning for image classification. *arXiv preprint arXiv:1905.00546* (2019).
[43] Ronghui You, Zihan Zhang, Ziye Wang, Suyang Dai, Hiroshi Mamitsuka, and Shanfeng Zhu. 2019. AttentionXML: Label tree-based attention-aware deep model for high-performance extreme multi-label text classification. In *Advances in Neural Information Processing Systems*, Vol. 32. 5820–5830.
[44] Hsiang-Fu Yu, Kai Zhong, and Inderjit S Dhillon. 2020. PECOS: Prediction for enormous and correlated output spaces. *arXiv preprint arXiv:2010.05878* (2020).
[45] Jiong Zhang, Wei-cheng Chang, Hsiang-fu Yu, and Inderjit S Dhillon. 2021. Fast Multi-Resolution Transformer Fine-tuning for Extreme Multi-label Text Classification. In *Advances in Neural Information Processing Systems*.
[46] Barret Zoph, Golnaz Ghiasi, Tsung-Yi Lin, Yin Cui, Hanxiao Liu, Ekin D Cubuk, and Quoc V Le. 2020. Rethinking pre-training and self-training. *arXiv preprint arXiv:2006.06882* (2020).

A ADDITIONAL EXPERIMENTS ON FS-XMC

Table 6: Results of FS-XMC where the training subset covers 1% positive pairs from the whole set.

Method	Precision			Recall				
	@1	@5	@10	@1	@3	@5	@10	@100
LF-Amazon-131K								
XR-Linear	5.37	2.66	1.68	2.81	3.92	4.09	4.26	4.99
Astec	3.29	2.04	1.41	1.93	3.33	3.77	4.06	5.06
SiameseXML	7.14	3.74	2.41	4.22	6.17	6.55	6.95	8.09
ZestXML	12.91	11.31	8.91	7.20	17.69	22.51	28.27	42.40
SentBERT	15.08	11.81	9.06	8.38	18.42	22.89	28.62	46.38
MPNet	15.26	12.30	9.42	8.56	19.35	23.98	29.91	48.06
MACLR (ours)	**18.92**	**16.17**	**12.62**	**10.98**	**25.64**	**32.16**	**39.46**	**58.24**
LF-WikiSeeAlso-320K								
XR-Linear	6.97	3.43	2.31	3.74	5.02	5.44	5.84	6.87
Astec	5.58	3.35	2.48	3.22	5.43	6.51	7.95	11.76
SiameseXML	9.87	5.22	3.59	5.84	8.57	9.53	10.60	13.04
ZestXML	10.40	8.18	6.49	5.57	11.65	14.52	18.81	33.20
SentBERT	18.85	14.23	11.22	11.16	21.77	26.94	33.78	55.88
MPNet	18.04	13.27	10.44	10.51	19.99	24.62	30.86	52.52
MACLR (ours)	**20.49**	**15.50**	**12.24**	**12.34**	**23.88**	**29.43**	**36.76**	**59.82**

Table 7: Results of FS-XMC where the training subset covers 5% positive pairs from the whole set.

Method	Precision			Recall				
	@1	@5	@10	@1	@3	@5	@10	@100
LF-Amazon-131K								
XR-Linear	11.20	5.82	3.80	5.98	8.56	9.18	9.80	12.79
Astec	10.71	6.50	4.52	6.12	10.23	11.67	13.35	18.15
SiameseXML	11.88	8.72	5.93	8.50	13.68	15.23	16.80	20.28
ZestXML	12.86	11.28	8.91	7.10	17.62	22.43	28.42	49.41
SentBERT	16.94	13.59	10.52	9.55	21.23	26.55	33.14	51.81
MPNet	17.48	13.58	10.61	9.95	21.38	26.83	33.60	52.31
MACLR (ours)	**19.75**	**16.45**	**12.87**	**11.18**	**25.99**	**32.70**	**40.38**	**59.82**
LF-WikiSeeAlso-320K								
XR-Linear	13.13	6.88	4.70	7.00	9.64	10.54	11.49	14.20
Astec	15.61	8.73	6.23	8.77	13.17	15.02	17.36	24.30
SiameseXML	16.51	9.68	6.96	9.40	14.78	16.97	19.48	25.26
ZestXML	17.68	8.51	6.85	10.63	12.01	15.20	20.08	43.10
SentBERT	20.12	15.01	11.87	12.05	23.01	28.40	35.52	58.41
MPNet	19.88	14.90	11.76	11.85	22.75	27.96	35.03	57.26
MACLR (ours)	**21.80**	**16.61**	**13.12**	**13.27**	**25.74**	**31.59**	**39.25**	**62.13**

In this section, we present additional experimental results for the setting of FS-XMC on LF-Amazon-131K and LF-WikiSeeAlso-320K. Instead of sampling a few-shot subset by the label coverage ratio, we turn to sampling based on the pair ratio. Specifically, suppose a training set $\mathcal{D}_{\text{train}} = \{(x_i, y_i)\}$ has $|\mathcal{D}_{\text{train}}|$ positive pairs. Each time we randomly sample a small ratio of δ (1% or 5% in our paper) pairs from the total set to constitute the few-shot subset. Then each subset has $\delta|\mathcal{D}_{\text{train}}|$ pairs for fine-tuning. Detailed results of $\delta = 1\%$ and $\delta = 5\%$ are presented in Table 6 and 7 respectively. MACLR is still the best-performing method and outperforms all other baselines significantly in precision and recall.